Risk Analytics in Banking: Understanding Loan Default Tendencies Based On Consumer and Loan Attributes

Math 1130: Final Project Report

By: Jingwen Yu, Dazhi Li, Priya Gill, Manas Agrawal

December 5th, 2023

Overall Research Topic: Risk factors influencing loan default tendencies: An in-depth analysis of consumer attributes, loan Characteristics and previous loan history.

Subtopic 1- Consumer Attributes:

Introduction of Problem:

We will determine if certain consumer demographics (e.g., age, education level, employment type) have a higher likelihood of defaulting on their loans than others. To answer this question, we looked at their education level and income type, respectively.

Data Analysis Steps:

The analysis of consumer attributes and loan defaults directly highlights how specific consumer attributes are related to the likelihood of loan defaults. It also equips lending institutions with valuable tools and knowledge about which attributes should be monitored closely, ultimately leading to more effective loan risk management. Additionally, it provides insights to potential borrowers, by offering guidance on how to improve and enhance their chances of loan approval based on certain attributes.

First of all, we focus on the application data file, and we get the information about the education level of all the people. Fortunately, no one's information is NaN, which makes our data more accurate. We get the data for the NAME_EDUCATION_TYPE column, Secondary/secondary special and Higher education respectively, and then we use a bar chart to display them respectively (**1.a**). From this chart, we can find that the majority of those who have overdue loans are Secondary/secondary special, while only about 25% of those with Higher education.

In addition, in the application document, we had planned to study the types of work they do. In this survey, however, we found gaps in the type of work that some people do. So I decided to look into their income types. Their work types are currently divided into four categories: State servant and Commercial associate and Pensioner, working. Through the analysis, we find that 62 percent are working people, 14 percent are State servants, 12 percent are commercial associates, and 12 percent are Pensioners **(1.b).**

Main Results & Conclusion:

According to the above analysis, we can find that the lack of education level and income type has an impact on whether people defer their loans. People with higher education levels are less likely to roll over their loans. For those in the income categories, Commercial associate and Pensioner, they are less likely to defer loans.

Disadvantages/Concerns:

Although we have cleaned up the data, we are still missing some data. For example, the data we plan to analyze above is the type of work people do, but there are many people whose data is missing. As for the data of education level, I only selected the top 10 for statistics, which may lead to bias and inaccurate results.

Subtopic 2- Loan Characteristics:

Introduction of Problem:

We will find what are the most common types of loans taken by consumers. We will also answer if there are specific loan characteristics that correlate with higher default rates. For our hypothesis, the type of cash loan is the most common one that people use. Loan duration and higher annuity amount can cause higher default rate.

Data Analysis Steps:

To find this answer, we need to investigate 2 csv files. For application_data.csv, we need to extract the columns of contract type, credit amount and annuity amount. For previous_application.csv, we need to look at "DAYS_FIRST_DUE" and "DAYS_LAST_DUE". To find the first question, we will analyze the number of times the contract type appeared. In the beginning, we used histograms to visualize the number of times the contract types appeared. We find that in the recent data and the previous data, the number of cash loans is the highest, although in the past, the number of consumer loans was close to it. In addition, it is far more than other types, such as revolving loans. Then we made histograms to observe the credits amount by contract type. We first use the groupby and mean function to find the relationship between them. We calculated the default rate by using the number of “Target” to divide total data. The rate is 0.0807. Before that, we find some data has NaN value, since we use dropna function to exclude this value for the “Target” column. The scatter plot can show the change of default rate with different annuity amount and goods price. The new column loan duration from the previous data, last due minus first due. Also, can be made a scatter plot for it. At the end, we used heatmap to find the default rate across regions based on population.

Main Results & Conclusion:

With the data, we find the cash loans have the highest occurrence rate. Furthermore, in the charts we observed that “Target” of 1 indicates a late payment, equivalent to a default. In graphs, the lower the annuity amount, the higher probability of late payments. The effect of the price of goods on default is the same. Therefore, the amount of annuity affects the default rate. The shorter loan duration, the larger the annuity amount and the smaller the number of late payments. In the heatmap, the denser the number of people, the lower the number of late payments.

Disadvantages/Concerns:

Because the amount of data is too large, it is impossible to check the data corresponding to each condition too carefully. Although the value of NaN is cleaned up, it may have other unknown values or strange values. Some data appeared in the previous table but not in the recent table. Therefore, some analyses cannot be representative of recent data.

Subtopic 3- Previous Loan History:

Introduction of Problem:

One of the main questions that we are trying to answer within our analysis, is whether a consumer's past loan repayment behavior predicts their future loan default tendencies. We hypothesize that a consumer's past loan repayment behavior, as indicated by loan statuses, credit amount relationships, and contract types in previous loan applications, is predictive of their future loan default tendencies.

Data Analysis:

In our analysis of this subtopic, we solely extracted information and manipulated data from the “previous_application.csv” file. Luckily, there wasn’t much cleaning involved in regard to NaN values, all rows had a value for the NAME_CONTRACT_TYPE and NAME_CONTRACT_STATUS columns. Our analysis for this subtopic is mainly focused on these two variables, since they give very valuable information.

The first step in our analysis process was to visualize the distribution of loan status for all loans (**Appendix 3.a**). The percentage of refused loans was 17.76%, and the percentage of approved loans was 65.63%. After calculating these two percentages, we made a count plot to show the

distribution of previous loan statuses. In this plot, the x-axis represents the four different categories of loan status, and the y-axis represents the number of loans that fall within each loan status. The count plot illustrated that the chance of loan approval significantly outweighs the other statuses, with loan cancellations being slightly lower than refusals.

Next, we analyzed the relationship between the previous application and credit amounts. To clarify, AMT_APPLICATION refers to how much credit the client initially asked for on the previous application, whereas AMT_CREDIT is how much the client really receives during the approval process. To analyze the relationship, we created a scatter plot, where x-axis is application amount and y-axis is credit amount **(Appendix 3.b)**. From the plot, we can see a strong linear relation which indicates that for most cases, the approval process matches the credit amount specified by the consumer.

Additionally, we analyzed how the credit amount relates to the loan contract status. Before creating the graph, we had to conduct a cleaning process, by extracting rows where AMT_CREDIT or AMT_APPLICATION is not equal to 0, and saving this as the "cleaned_previous_application" data frame. We created a data_for_boxplot data frame, containing only the columns NAME_CONTRACT_STATUS and AMT_CREDIT from the cleaned dataset. We created a box plot (**Appendix 3.c**), by adjusting the y-scale to the logarithmic function, in order to better visualize a wide range of credit amounts. The x-axis on the graph represents the categories for contract status.

Lastly, we analyzed the correlation between the loan status, and all three of the contract types (cash loan, revolving loan, and consumer loan), by creating a grouped bar chart **(Appendix 3.d)**. We represented the number of loans on the y-axis, the contract types on the x-axis, and coloured the bars for each contract type based on the contract status (shown in legend). From this graph, we can evidently see that consumer loans are a lot more likely to get approved than any other type of loan. We can also see that cash loans have a very low approval rate, compared to the number of cash loans that are canceled and refused.

Main Results & Conclusion:

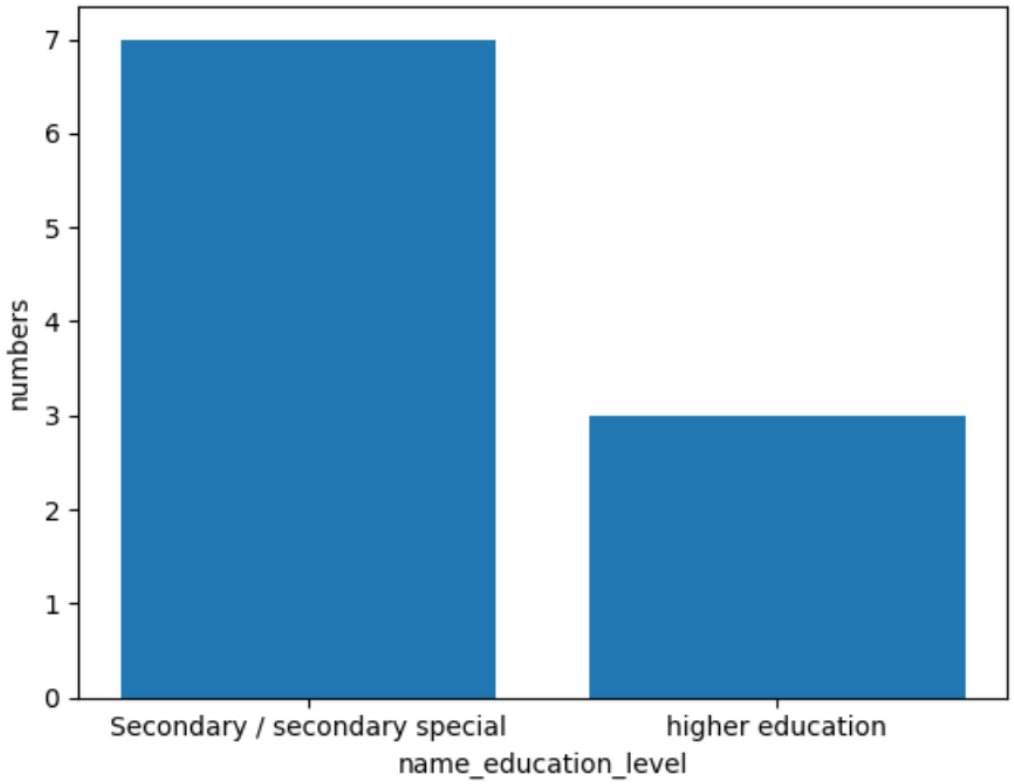
The analysis that we conducted on the previous loan application data strongly supports the hypothesis that past loan repayment behavior can predict future loan default tendencies. From the visualization of loan status distribution, it has been shown that there is a notably higher loan approval chance compared to other statuses, which potentially contributes to the recent rise in default rates. We also learned that there is a strong linear relationship between the application and credit amounts implies that, in most cases, the approved credit amount aligns closely with what consumers initially requested. When observing the analysis of credit amounts concerning contract status, the consumer loans exhibited higher approval rates compared to other types, while cash loans displayed significantly lower approval rates. This emphasizes that there is potential for default tendencies in cash loans, which is why the approval rate is much lower. In summary, the analysis strongly supports the hypothesis, showcasing that certain loan statuses, credit amount relationships, and contract types significantly influence future loan default tendencies.

Disadvantages/Concerns:

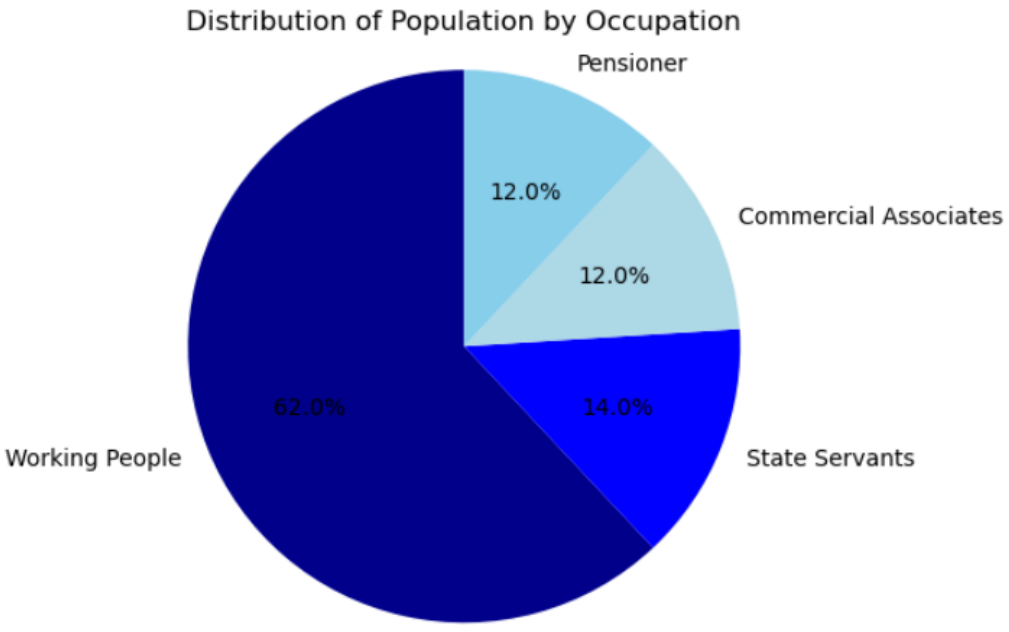
There are some disadvantages and concerns that come with performing a data analysis on the given data set. Specifically, there are limited variables in the pervious_application.csv file. If this file had a "TARGET" variable, we could have done much deeper analysis, however since it did not, we primarily focused on the contract type and status instead. We also have to remember that even though our analysis demonstrates strong correlations between past loan behaviors and future defaults, it's essential to consider the fact that correlation doesn't always imply causation. To be more specific, other external and uncontrollable factors that were not accounted for in this analysis, might influence default tendencies as well.

Appendix

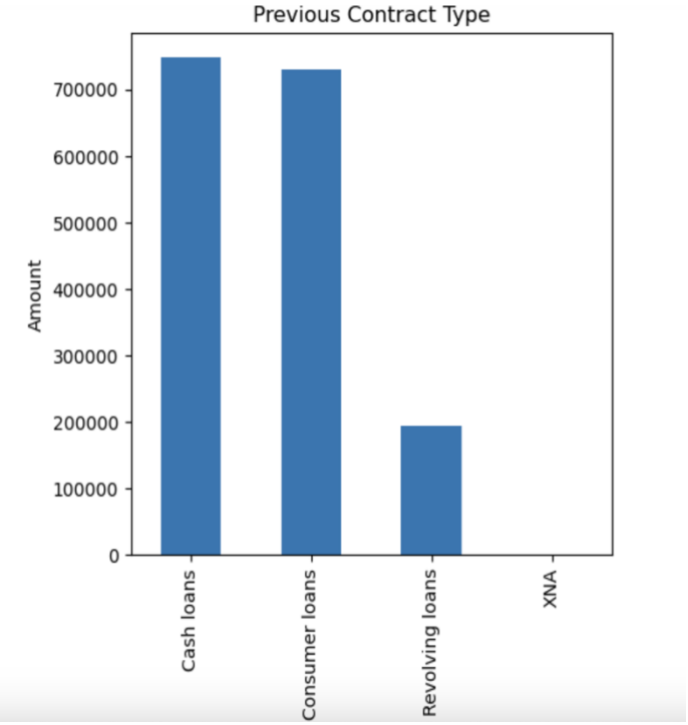
1.a



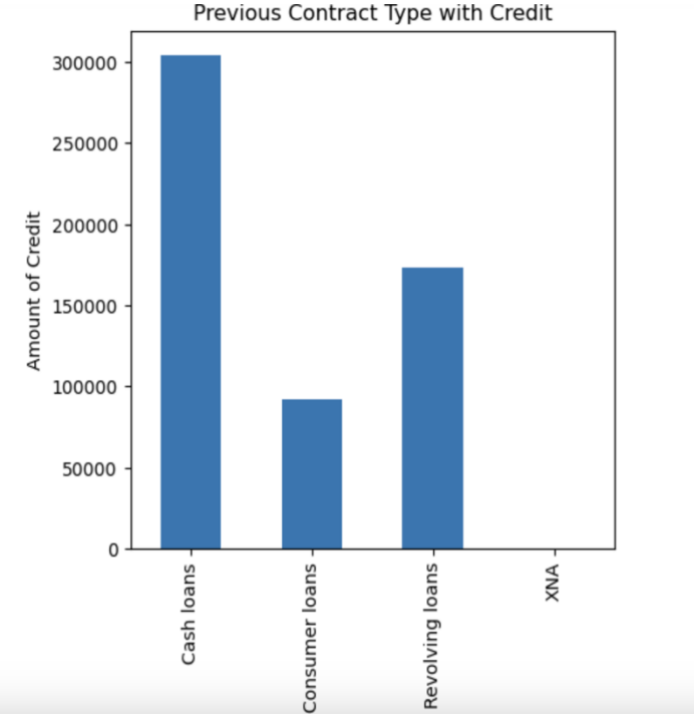
1.b



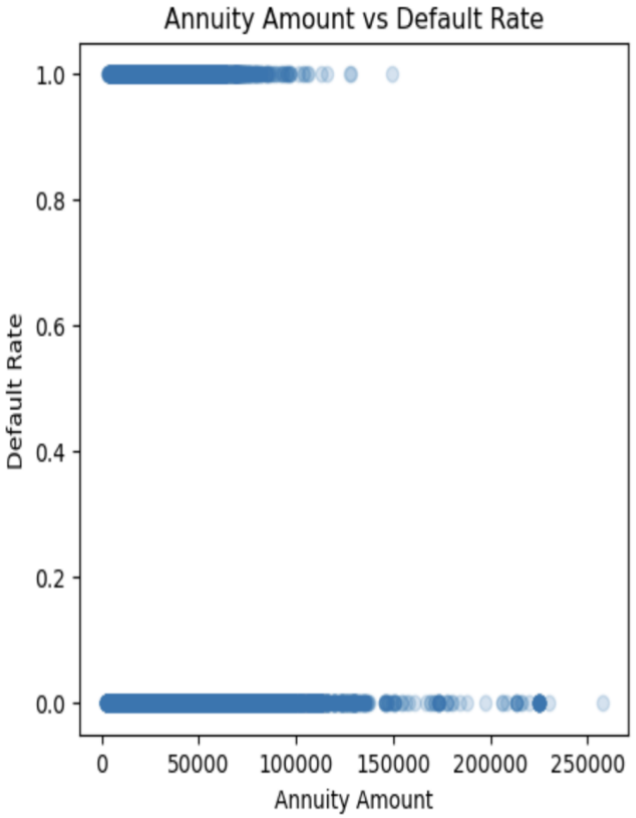
2.a


Previous Contract Type
Amount
0
100000
200000
300000
400000
500000
600000
700000
Cash loans
Consumer loans
Revolving loans
XNA

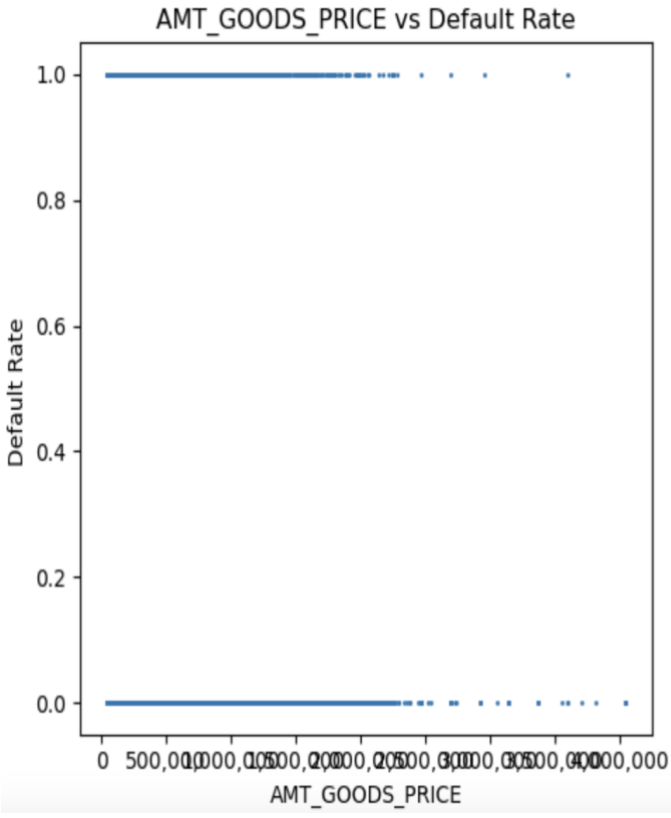
2.b


Previous Contract Type with Credit
Amount of Credit
0
50000
100000
150000
200000
250000
300000
Cash loans
Consumer loans
Revolving loans
XNA

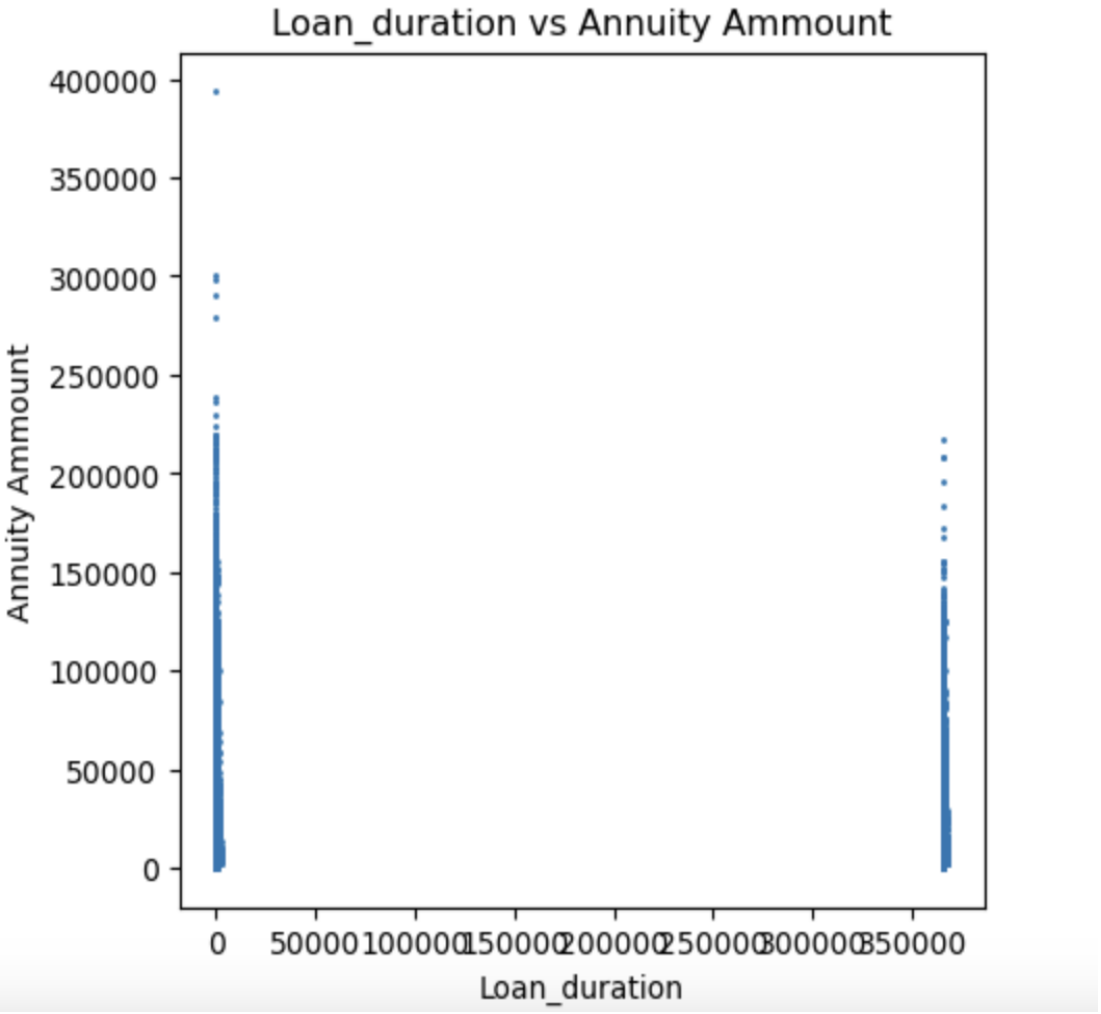
2.d


Annuity Amount vs Default Rate
Default Rate
0.0
0.2
0.4
0.6
0.8
1.0
0
50000
100000
150000
200000
250000
Annuity Amount

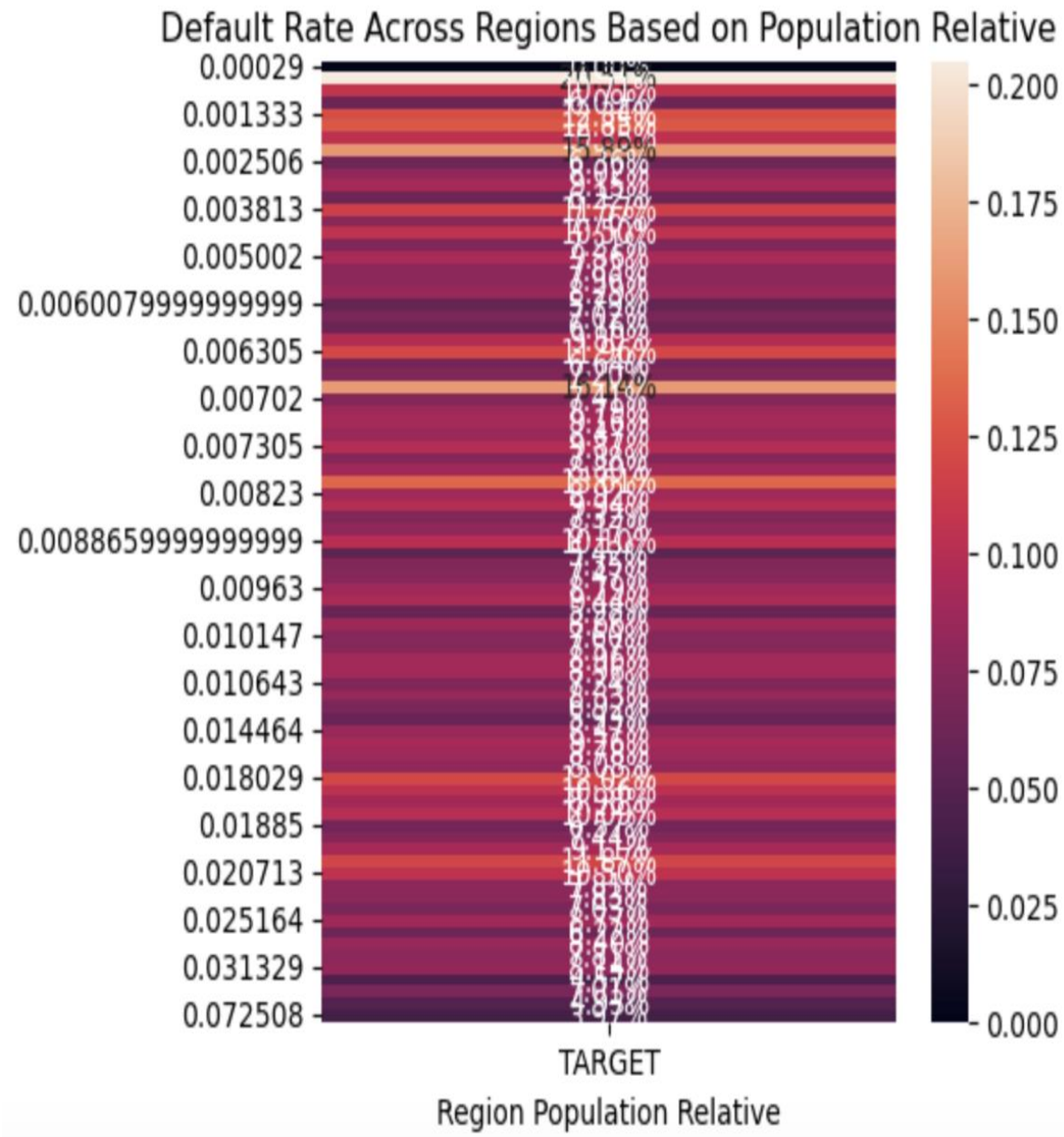
2.e


AMT_GOODS_PRICE vs Default Rate
Default Rate
0.0
0.2
0.4
0.6
0.8
1.0
AMT_GOODS_PRICE

2.f

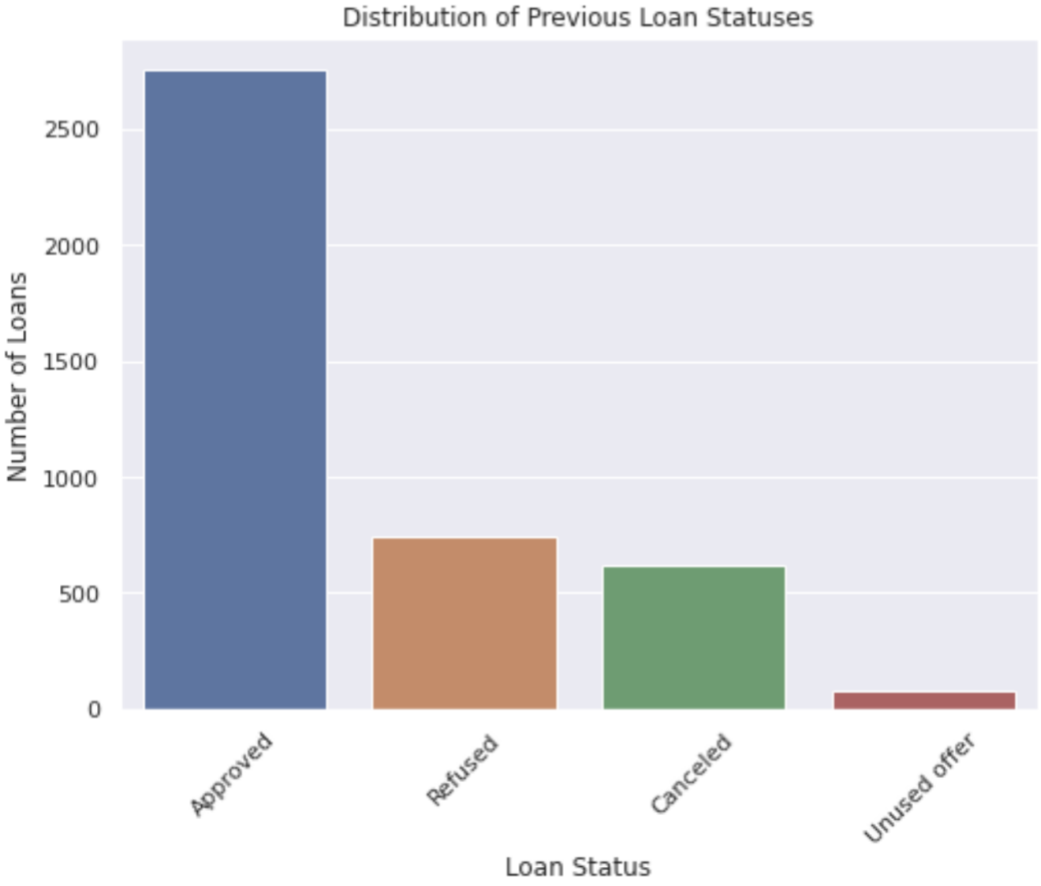

Loan_duration vs Annuity Ammount
Annuity Ammount
400000
350000
300000
250000
200000
150000
100000
50000
0
0
50000
100000
150000
200000
250000
300000
350000
Loan_duration

2.g

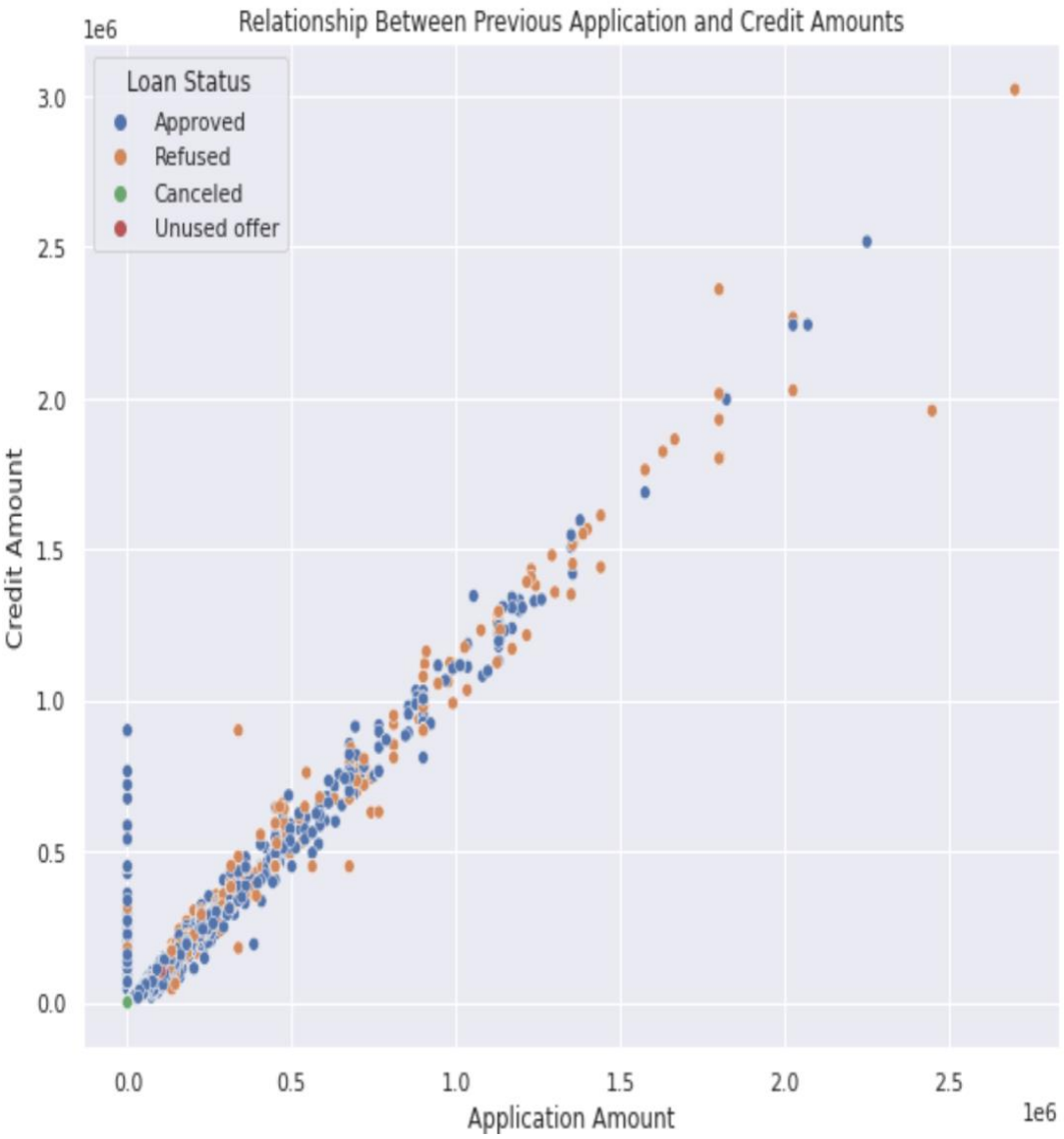

Default Rate Across Regions Based on Population Relative
0.00029
0.001333
0.002506
0.003813
0.005002
0.0060079999999999
0.006305
0.00702
0.007305
0.00823
0.0088659999999999
0.00963
0.010147
0.010643
0.014464
0.018029
0.01885
0.020713
0.025164
0.031329
0.072508
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
TARGET
Region Population Relative

3. a

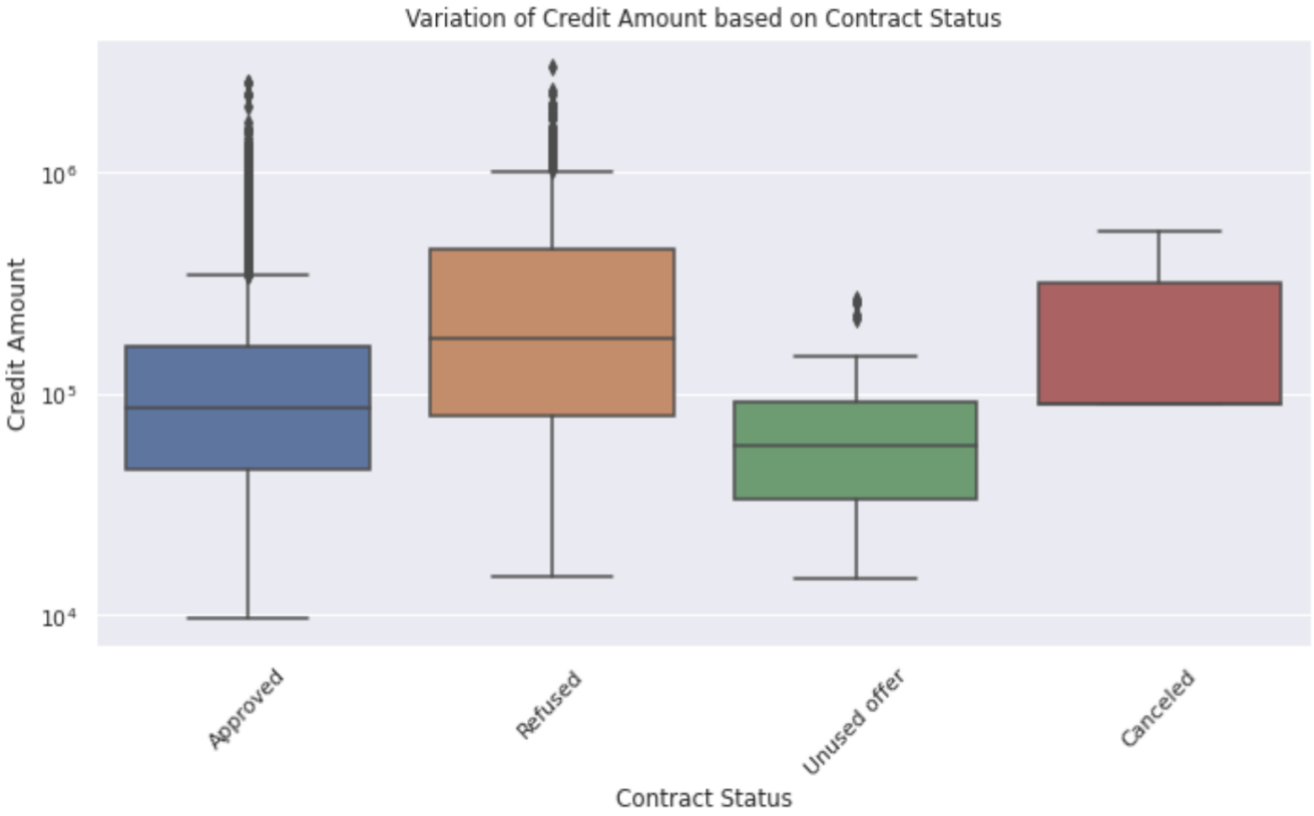
```
Previous Loan Refusal Rate: 17.76%
Previous Loan Approval Rate: 65.63%
```



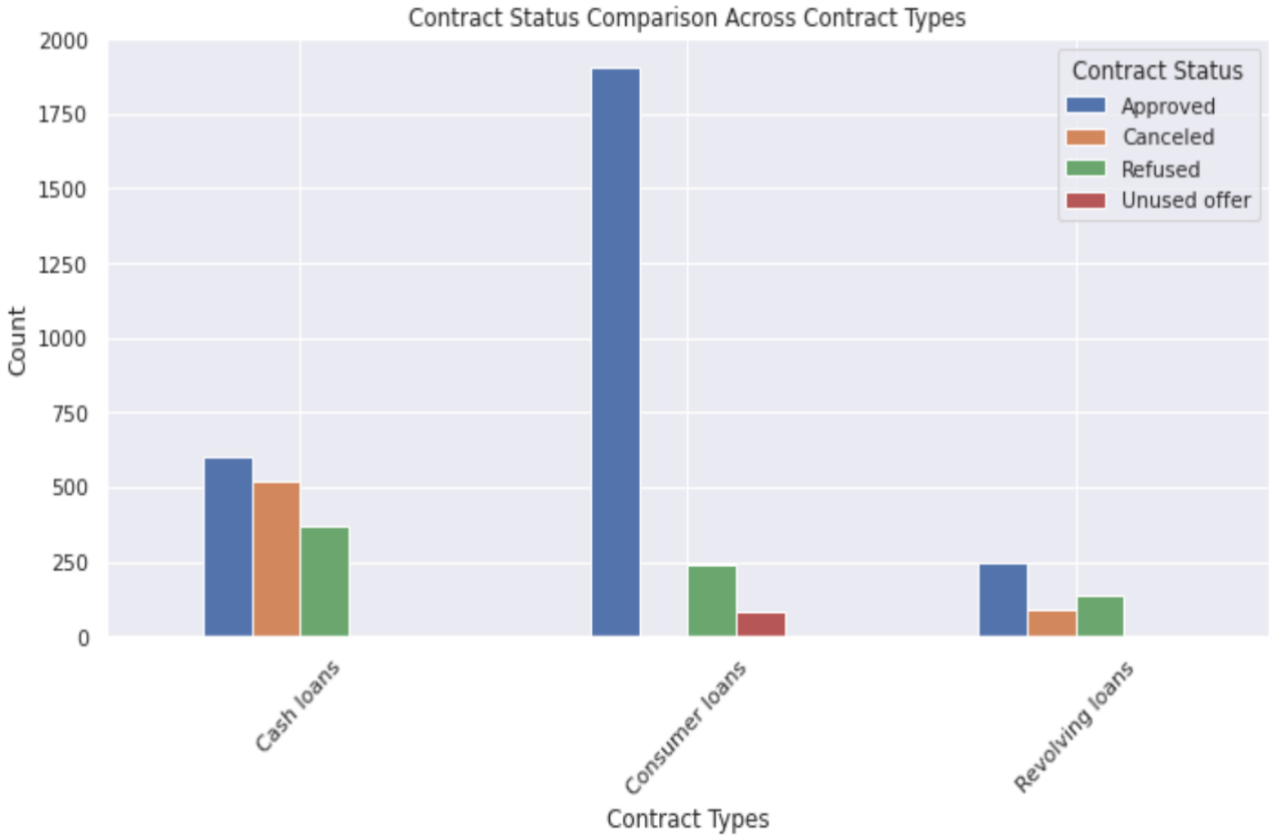
3. b


Relationship Between Previous Application and Credit Amounts
1e6
Loan Status
Approved
Refused
Canceled
Unused offer
Credit Amount
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
2.5
Application Amount
1e6

3. c


Variation of Credit Amount based on Contract Status
Credit Amount
10^6
10^5
10^4
Approved
Refused
Unused offer
Canceled
Contract Status

3. d


Contract Status Comparison Across Contract Types
Contract Status
Approved
Canceled
Refused
Unused offer
Count
2000
1750
1500
1250
1000
750
500
250
0
Cash loans
Consumer loans
Revolving loans
Contract Types